Modeling the Interaction between Speech and Gesture

MS-CIS-94-23
LINC LAB 268
HUMAN MODELING & SIMULATION LAB 61

Justine Cassell
Matthew Stone
Brett Douville
Scott Prevost
Brett Achorn
Mark Steedman
Norm Badler
Catherine Pelachaud

DTIC
ELECTE
FEB 08 1995
S G D

19950203 211

University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

May 1994

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

MASTER COPY   KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
technical report

4. TITLE AND SUBTITLE
Modeling the Interaction between Speech and Gesture

5. FUNDING NUMBERS
DAAL03-89-C-0031

6. AUTHOR(S)
J. Cassell, M. Stone, B. Douville, S. Prevost, B. Achorn
M. Steedman, N. Badler, C. Pelachaud

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer and Information Science Department
University of Pennsylvania
200 S. 33rd Street
Philadelphia, PA 19104-6389

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 26779.39-MA-AI

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This paper describes an implemented system that generates spoken dialogue, including speech, intonation, and gesture, using two copies of an identical program that differ only in knowledge of the world and which must cooperate to accomplish a goal. The output of the dialogue generation is used to drive a three-dimensional interactive animated model – two graphic figures on a computer screen who speak and gesture according to the rules of the system. The system is based upon a formal, predictive and explanatory theory of the gesture-speech relationship. A felicitous outcome is a working system to realize autonomous animated conversational agents for virtual reality and other purposes, and a tool for investigating the relationship between speech and gesture.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

Modeling the Interaction between Speech and Gesture

Justine Cassell **Matthew Stone** **Brett Douville** **Scott Prevost**
Brett Achorn **Mark Steedman** **Norm Badler** **Catherine Pelachaud**

Computer & Information Science
University of Pennsylvania
Philadelphia, PA 19104-6389

Abstract

This paper describes an implemented system that generates spoken dialogue, including speech, intonation, and gesture, using two copies of an identical program that differ only in knowledge of the world and which must cooperate to accomplish a goal. The output of the dialogue generation is used to drive a three-dimensional interactive animated model – two graphic figures on a computer screen who speak and gesture according to the rules of the system. The system is based upon a formal, predictive and explanatory theory of the gesture-speech relationship. A felicitous outcome is a working system to realize autonomous animated conversational agents for virtual reality and other purposes, and a tool for investigating the relationship between speech and gesture.

Accesion For
NTIS CRA&I ☒
DTIC TAB □
Unannounced □
Justification ________
By ________
Distribution /
Availability Codes
Dist | Avail and / or Special
A-1

To appear in the proceedings of Cognitive Science Society Annual Meeting 1994

1 Introduction

In recent work (e.g. Alibali and Goldin-Meadow, 1993; Cassell et al., 1993; McNeill, 1992) it has been argued that spontaneous gesture produced unwittingly by speakers and the speech it accompanies form an integrated conceptual system. Thus, gesture is not a *translation* of speech, or irrelevant to speech. Gesture and speech are different communicative manifestations of one single mental representation. However, until now research on the relationship between gesture and speech has been difficult to evaluate because of its descriptive nature. One way to move from descriptive to predictive theories is via formal models, which point up gaps in knowledge and fuzziness in theoretical explanations. We present such a model, embodied in a dialogue generation program that drives two animated human figures, simulating conversational interaction.

The dialogue generation is a novel generalization of earlier work by Power, 1977, Houghton and Pearson, 1988, and extends earlier work by Prevost and Steedman, 1993a, to include further distinctions of discourse information, allowing gesture and more sophisticated conversational intonation to be generated along with speech. We believe that no computational system that automatically generates conversations between two autonomous human-like agents, with appropriate and synchronized speech, intonation and hand gestures, has been implemented before. In the remainder of the introduction we describe research on the relationship between gesture and speech that underlies our attempt to simulate the behavior. We follow that with a discussion of intonation and information structure, and give a set of rules for gesture generation with respect to those two linguistic variables. We then describe the various modules of the simulation: the dialogue generation program, speech and intonation synthesis, gesture integration, and animation interface.

Four basic types of gestures occur only during speaking (McNeill, 1992); these four types of speech-associated gesture have been the focus of the majority of research on the cognitive basis of the gesture-speech relationship, including our own. *Iconics* represent some feature of the accompanying speech, such as sketching a small rectangular space with one's two hands while saying "Did you bring your CHECKBOOK?". *Metaphorics* represent an abstract feature concurrently spoken about, such as forming a jaw-like shape with one hand, and pulling it towards one's body while saying "You must WITHDRAW money.". *Deictics* indicate a point in space. They accompany reference to persons, places and other spatializeable discourse entities. An example is pointing to the ground while saying "Do you have an account at Mellon or at THIS bank?". Finally, *Beats* are small formless waves of the hand that occur with heavily emphasized words, occasions of turning over the floor to another speaker, and other kinds of special linguistic work. An example is waving one's left hand briefly up and down along with the stressed words in the phrase "Go AHEAD."

Evidence from many sources suggests a close relationship between speech and gesture. At the prosodic level, Kendon, 1974 found that the stroke phase (most effortful part) of these gestures tends to co-occur with or just before the phonologically most prominent syllable of the accompanying speech. At a cognitive level, Cassell et al., 1993 established that listeners rely on information conveyed in gesture as they try to comprehend a story; Alibali and Goldin-Meadow, 1993 showed that children may express in gesture information that they cannot yet express in speech. Other evidence comes from the sheer frequency of gestures during speech. About three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another (McNeill, 1992), and perhaps surprisingly, although the proportion of gesture types may change, all of these gestures, and spontaneous gesturing in general, are found in discourses by speakers of most languages.

In this paper our primary concern is with the semantic and pragmatic relationship between the two media. Gesture and speech do not always manifest the same information about an idea, but for adults what they convey is always complementary. That is, gesture may depict the way in which an action was carried out when this aspect of meaning is not depicted in speech. It has been suggested (Kendon, 1994) that those concepts difficult to express in language may be conveyed by gesture. Thus simultaneity of two events, or the respective locations of two objects may be expressed by the position of the two hands. In this sense, the gesture-speech relationship resembles the interaction of words and graphics in the generation of multimodal text (Feiner and McKeown, 1991; Wahlster et al., 1991). In storytelling, narrative structure may be indexed by differential use of gesture: iconic gestures tend to occur with plot-advancing description of the action, deictic gestures with the introduction of new characters, and beat gestures at the boundaries of episodes (Cassell and McNeill, 1991). Until now, however, there has been no attempt to predict when and what kinds of gestures will occur in a discourse.

We propose to use the level of *information structure* to capture the regularities of gesture occurrence. The information structure of an utterance defines its relation to other utterances in a discourse and to propositions in the relevant knowledge pool. Although a sentence like "George withdrew fifty dollars" has a clear semantic interpretation, the semantics does not indicate how the proposition relates to other propositions in the discourse. For example, the sentence might be an equally appropriate response to the questions "Who withdrew fifty dollars," "What did George withdraw," "What did George do," or even "What happened." Which question is asked determines which items in the response are most important or salient, which in turn determines how the phrase is uttered. These types of salience distinctions are encoded in the information structure representation of an utterance.

Following Halliday and others (Halliday, 1967; Hajičová and Sgall, 1988), we use the terms *theme* and *rheme* to denote two distinct information structural attributes of an utterance.[1] The theme roughly corresponds to what the utterance is about. The rheme corresponds to what the speaker has to contribute concerning the theme. Depending on the discourse context, a given utterance may be divided on semantic and pragmatic grounds into thematic and rhematic constituents in a variety of ways. For example, given the utterance "George withdrew fifty dollars," we might consider the theme to be 'How much money George withdrew' and the rheme to be 'fifty dollars.'

[1] Although note that we drop Halliday's assumption that themes occur only in sentence–initial position. Functionally similar distinctions in this context are *topic/comment*, *given/new*, and the scale of *communicative dynamism*.

Within information structural constituents, we define the semantic interpretations of certain items as being either *focused* or *background*. Items may be focused for a variety of reasons, including emphasizing their newness in the discourse or making contrastive distinctions among salient discourse entities. For example, in a theme concerning 'How much money George withdrew' we may say that 'George' may be the focus because it stands in contrast to some other salient discourse entity, say 'Gilbert'. We also mark the representation of entities in information structure with their status in the discourse. Entities are considered either new to discourse and hearer (indefinites), new to discourse but not to hearer (definites on first mention), or old (all others) (Prince, 1992).

Distinct intonational tunes have been shown to be associated with the thematic and rhematic parts of an utterance for certain classes of dialogue (Prevost and Steedman, 1993a; Prevost and Steedman, 1993b; Steedman, 1991). In particular, we note that the standard rise-fall intonation generally occurs with the rhematic part of many types of utterances. The rise-fall intonation is realized as a pitch peak on the primary-stress syllable of the focused word, followed by an immediate fall to a lower pitch which is then sustained for the duration of the phrase. The rhematic part of yes/no interrogatives is often accompanied by a fall-rise intonation, realized as a low pitch target on the primary-stress syllable of the focused word, followed by an immediate rise to a sustained higher pitch. Thematic elements of an utterance are often marked by a rise-fall-rise intonation, realized by a rise to a high pitch target on the primary-stress syllable, followed by an immediate fall to a lower pitch with another pitch rise occurring at the end of the phrase. The following examples illustrate the coupling of tunes with themes and rhemes.

(1)

Q: I know who withdrew three dollars,
but who withdrew fifty dollars?

A: (GEORGE)$_{\text{rheme}}$ (withdrew FIFTY dollars)$_{\text{theme}}$

(2)

Q: I know how many dollars Gilbert withdrew,
but how many did George withdraw?

A: (GEORGE withdrew)$_{\text{theme}}$ (FIFTY dollars)$_{\text{rheme}}$

We claim that that, just as intonational features are associated with information structural aspects of discourse, so too are gestures. We propose the following rules to predict the location and types of gestures as a function of information structure:

- Non-beat gestures accompany verb phrases in the rheme, and hearer new references, as follows: words with literally spatial content get iconic gestures; those with metaphorically spatial content get appropriate metaphorics; words with spatializable content get deictics.

- Beat gestures are generated for verb phrases in the rheme and for hearer new references when the semantic content cannot be represented spatially.

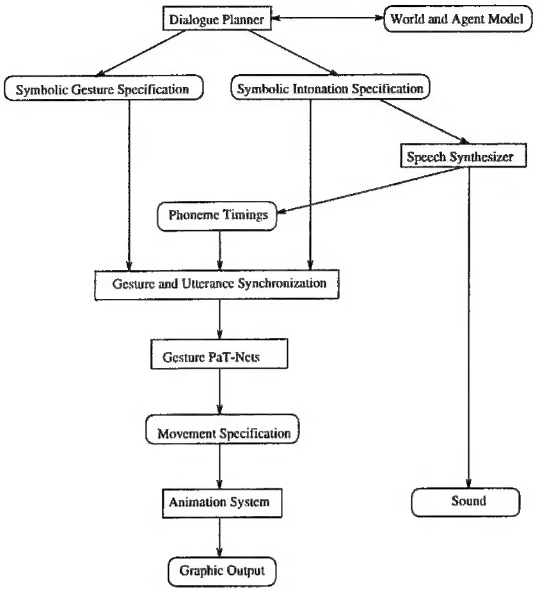

Figure 1: Interaction of components

- Beats accompany discourse new definite references.

Generated gestures associated with a word are aligned with the stressed syllable of that word. This is a straightforward task for beat gestures which are simply waves of the hand. In contrast, the other gestures have a preparation phase which occurs before the stroke; accordingly, gestures with a preparation phase must start at the beginning of the intonational phrase in which the associated word occurs to ensure that the stroke can occur on that word.

2 Implementation

The objective of this project is to provide a testbed where predictive accounts of gesture, such as the rules given above, can be formalized and evaluated. This goal places certain demands on the generation of content for the "computational stage" (McNeill, 1992) shared by speech and gesture. In particular, the generation process must provide precise and explicit representations of the concepts, such as information structure, to which the theory of gesture refers.

The solution adopted here is to simulate the world and discourse actions of an agent interacting with another agent in the service of accomplishing a goal in a simple environment. For the present implementation, the 'bank domain' was chosen because in it there are two agents who must interact linguistically to accomplish a goal. The complexity of the domain and the steps followed by the agents are analogous to those of Power, 1977 and Houghton and Pearson, 1988, but here the model is enriched with explicit representations of the structure of discourse and the relationship of the structure to the agents' domain plans. The following describes the dialogue generation system in this light. We then turn to the other elements of the system represented in Figure 1.

The selection of content for the dialogue in our system is performed by two cascaded planners. The first is the domain planner, which manages the plans governing the concrete actions which the agents will execute; the second is the discourse planner, which manages the communicative actions the agents must take in order to agree on a domain plan and in order to remain synchronized while executing a domain plan.

The input to the domain planner is a database of facts describing the way the world works, the goals of the agents, and the beliefs of the agents about the world, including the beliefs of the agents about each other. The domain planner executes by decomposing an agent's current goals into a series of more specific goals according to the hierarchical relationship between actions specified in the agent's beliefs about the world. An agent's goals may be of one of two forms: to obtain some piece of information, or to ensure that some state holds in the world; questions can be used to achieve either kind of goal, but planning decompositions are only appropriate for the second kind. Once decomposition resolves a plan into a sequence of actions to performed, the domain planner causes the agents to execute those actions in sequence. As these goal expansions and action executions take place, the domain planner also dictates discourse goals that agents must adopt in order to maintain and exploit cooperation with their conversational partner.

The domain planner transmits its instructions to take communicative actions to the discourse planner by suspending operation when such instructions are generated and relinquishing control to the discourse planner. Several stages of processing and conversational interaction may occur before these discourse goals are achieved. The discourse planner must identify how the goal submitted by the domain planner relates to other discourse goals that may still be in progress. Then content for a particular utterance is selected on the basis of how the discourse goal is decomposed into sequences of actions that might achieve it. The following fragment of dialogue was produced by the dialogue planner. The previous segment of discourse established that Gilbert is a bank teller, and George a customer, and that George has asked Gilbert for help in obtaining $50.

The dialogue is repetitive and explicit in its goals exactly because the two agents have to specify in advance each of the goals they are working towards, and steps they are following. The dialogue generation program has none of the conversational inferences that allow humans to follow leaps of reasoning. True conversational inference greatly complicates the conversational planning component of dialogue generation. Moreover, the explicit expression of all conversational moves can be viewed as a trace of the steps that more sophisticated conversational inferencers would have to follow.

Gilbert: Do you have a blank check?
George: Yes, I have a blank check.
Gilbert: Do you have an account for the check?
George: Yes, I have an account for the check.
Gilbert: Does the account contain at least fifty dollars?
George: Yes, the account contains eighty dollars.
Gilbert: Get the check made out to you for fifty dollars and then I can withdraw fifty dollars for you.
George: All right, let's get the check made out to me for fifty dollars.

The domain planner and the discourse planner offer a number of explicit representational structures which could serve as input in formulating rules of gesture and intonation. At any point, of course, each agent has a representation of the domain plan that is being executed, and of the constituents of discourse that go into discussion of the plan. Explicit links between these two structures indicate what part of the plan each discourse segment concerns; these links ensure that conversation is coordinated and understood. These three kinds of information form the basis for two additional levels of representation, which are maintained solely for their possible relevance to linguistic processes. First, a model of attention (the attentional state) indicates which entities are known to the participants, which entities have been referred to, and how salient those entities are. The attentional state for some utterance in the discourse consists of a list that contains, for each discourse segment which dominates that utterance, the sets of entities mentioned in that segment. These sets are ordered so that the entities referred to in larger segments are less salient than the entities referred to in segments they dominate. Second, a record of the purposes generated by the planner which initiated discourse actions is kept. Of course, it may happen that only the agent who initiated an action knows this purpose exactly. Accordingly, both parties also separately record the most specific purpose for a segment for which evidence has been given. This architecture of intentional structure, attentional state, and discourse purposes, and the relationship between them was first proposed by Grosz and Sidner, 1986; the implementation of these notions here follows their suggestions as closely as possible. We use these representations to reconstruct the information structure of the dialogue as follows.

- Material is classified as thematic if it occurs in some part of the speaker's discourse purpose in the current constituent or its ancestors for which evidence has been given.
- Material is classified as rhematic if it occurs only in that part of the speaker's discourse purpose in the current segment or its ancestors for which evidence has not been provided.
- Information not meeting either of these criteria constitutes linguistic formulae, which are irrelevant to the speaker's purpose, and are also labelled as thematic.

Focus is assigned to references according to the theory of contrast in Prevost and Steedman, 1993a, while the discourse status of entities is determined from the agents' knowledge of each other and from the attentional state. Finally. the semantic class of constituents is retrieved from a dictionary associating semantic representations with possible gestures that might represent them.[2]

These structures permit application of the rules for generation of gestures and intonation given above. A variant of Prevost and Steedman's algorithm is used to do this, thus generating English text annotated with intonational cues and gestural instructions from information structures. These intonational and gesture features are attached to words in the dialogue and may alternatively be interpreted as occurring at the start of the associated word, on the stressed syllable of the word, or at the end of the word, depending on the feature. In order for the gestures to appear at the proper times in the animation, the two streams must be synchronized with the synthesized speech.

The intonation stream provides an abstract representation which is automatically translated to a form suitable for input to the speech synthesis component. We currently use the AT&T Bell Laboratories TTS synthesizer to produce the actual speech wave and phoneme timings (Liberman and Buchsbaum, 1985). The following example demonstrates the intonational tunes specified by the dialogue planner and sent to the speech synthesizer.

Get the	CHECK	made	OUT	to you for fifty	dollars
	rise		rise		fall
and	THEN	I can	WITHDRAW	fifty dollars for	you.
	rise-fall-rise		rise		fall

After transforming the utterances into proper input for the synthesizer and generating the speech wave and phoneme timings, the durational outputs from the synthesis are merged by rule with the abstract intonational and gestural notations. This detailed timing information (to the centisecond) allows synchronization of the gestural animations with the speech, as described below.

3 Gesture Integration and Animation

In the research presented here the interaction between speech and gesture is modeled in such a form that it can drive an animation system[3]. The input to the animation system does not specify every small movement of the hands, in order to create a flexible system that separates issues of semantics from physiology. The system does, however, take

[2] This solution is provisional: a richer semantics would include the features relevant for gesture generation, so that the form of gestures could be generated algorithmically from the semantics. Note also, however, that following Kendon, 1994 we are led to believe that gestures may be more standardized than previously thought.

[3] Another model currently in progress generates gaze and head movements, and synchronizes gestures with these facial parameters as well as with movements of the lips (Pelachaud et al., 1991; Cassell et al., 1994).

into account temporal deformations of gestures due to the demands of synchronizing gestures with speech and with one another.

Gesture production is carried out by a group of Parallel Transition Networks (PaT-Nets), finite state machines several of which can be run in tandem (Becket, 1994). PaT-Nets govern three processes, two of which concern the direct production of gesture through the animation system. The first, parse-net, is a control network which parses the output of the speech synthesis module described above. This finite state machine parses phoneme representations one utterance at a time; in the current domain, this means also that one speaker turn is parsed at a time.

Upon the signalling of a particular gesture, parse-net will instantiate one of two additional PaT-Nets; if the gesture is a beat, the finite state machine representing beats ("beat-net") will be called, and if a deictic, iconic, or metaphoric, the network representing these types of gestures ("gest-net") will be called. This separation is motivated by the "rhythm hypothesis" (Tuite, 1993) which posits that beats arise from the underlying rhythmical pulse of speaking, while other gestures arise from meaning representations. In addition, beats are often found superimposed over the other types of gestures, and such a separation facilitates implementation of superposition. Finally, since one of the goals of the model is to reflect differences in behavior among gesture types, this system provides for control of freedom versus boundedness in gestures (e.g. an iconic gesture or emblem is tightly constrained to a particular standard of well-formedness, while beats display free movement); free gestures may most easily be generated by a separate PaT-Net whose parameters include this feature.

Gesture and beat finite state machines are built as necessary by the parser, so that the gestures can be represented as they arise. The newly created instances of the gesture and beat PaT-Nets do not exit immediately upon creating their respective gestures; rather, they pause and await further commands from the calling network, in this case, parse-net. This is to allow for the phenomenon of gesture coarticulation, in which two gestures may occur in an utterance without intermediary relaxation, i.e. without dropping the hands or, in some cases, without relaxing handshape. Once the end of the current utterance is reached, the parser adds another level of control: it forces exit without relaxation of all gestures except the gesture at the top of the stack; this final gesture is followed by a relaxation of the arms, hands, and wrists.

The animation itself is carried out by *Jack*®, a program for controlling articulated objects, especially human figures. The figures have joints and behaviors designed to generate realistic motion. Additional modules can be added to deal with new domains, such as gesture.

The PaT-Net system issues gesture requests to the animation system, telling the figure to either rest, make a beat motion, or make a gesture involving the hand, wrist, and/or arm. Four motion modules have been added to the *Jack* system: hand motion, wrist motion, arm motion, and beat motion, each which may be specified separately for each side of the body. The animation system isolates the higher level PaT-Net system from the details of the human figure geometry, biomechanical modeling, and joint control functions.

The hand motions can be specified in terms of an expandable library of hand

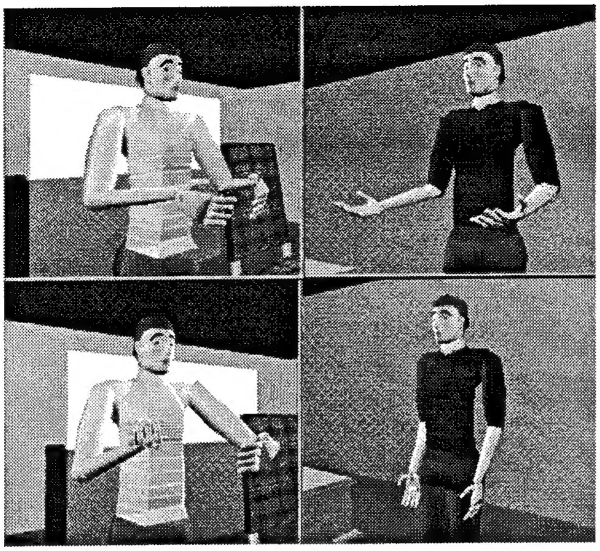

Figure 2: Examples of symbolic gesture specification

shapes, and the current system is based on the American Sign Language alphabet. An additional parameter controls the laxness of the handshape. The animation system moves the fingers from one position to another, attempting to get as close to the goal positions as possible within the contraints of the time alloted and the velocity limits of the finger joints. The result is that as the speed of the gesture increases, the gestures will 'coarticulate' in a realistic manner.

The wrist position goals are specified in terms of the hand direction relative to the figure (e.g. fingers forward and palm up). The animation system automatically limits the wrist to a realistic range of motion. Beat motions are a specialized form of wrist motion. Rather than having the goal specified, the goal is automatically generated based on the current position of the wrist. The animation system selects the most comfortable way for the figure to gesture in that situation and moves the wrist accordingly. The arm motions are specified in a manner similar to that of the wrists, except that relative spatial positions (e.g. near to the body, far left, and chest high) are given instead of orientations.

4 Gesture Output

In Figure 2, we see examples of how gestures are generated from the discourse content illustrated by the fragment of dialogue reproduced above.

1. "Do you have a BLANK CHECK?"
   - In the first frame, an iconic gesture (representing a rectangular check) is generated from the first mention (new to hearer) of the entity 'blank check'.
2. "Will you HELP me get fifty dollars?"
   - In the second frame, a metaphoric gesture (the common *propose* gesture, representing the request for help as a proposal that can be offered to the listener) is generated because of the first mention (new to hearer) of the request for help.
3. "You can WRITE the check."
   - In the third frame, an iconic gesture (representing writing on a piece of paper) is generated from the first mention of the concrete action of 'writing a check'.
4. "I will WAIT for you to withdraw fifty dollars for me."
   - In the fourth frame, a beat gesture (a movement of the hand up and down) is generated from the first mention of the concept 'wait for', which cannot be represented spatially.

5 Conclusion

Most research on gesture has been descriptive and distributional. With this descriptive infrastructure in place, it is now possible to attempt formal and predictive theories of gesture use. The theory of gesture-speech interaction described above allowed us to specify rules and write algorithms that drive an animated model of verbal and non-verbal behaviors in conversational interaction. The shortcomings of the implementation are as interesting as its successes. In particular, while the theory quite successfully specified when gestures might be expected in a discourse, and what the temporal relationship between those gestures and speech is, it lacked a basis for distributing communicative load among gesture, speech content, and intonation. That is, the discourse model might generate turn-taking phrases such as "Go ahead" and also generate beat gestures to accompany those phrases. In natural human interaction, it is more likely that either gesture or speech take on such a linguistic function, but not the two systems simultaneously. We expect to return to the parallel established above with automatically generated coordinated multimedia presentations as a way of improving our model. In the meantime, we have demonstrated that gesture and speech can be generated from one single underlying semantic representation, reflecting an integrated conceptual system. And in so doing, we have implemented autonomous animated conversational agents, and a testbed to explore further theories of gesture.

6 Acknowledgements

This research is partially supported by NSF VPW GER-9350179; NSF IRI91-17110; NSF IRI90-18513; NSF graduate fellowships; ARO Grant DAAL03-89-C-0031 including participation by the U.S. Army Research Laboratory (Aberdeen); U.S. Air Force DEPTH contract through Hughes Missile Systems F33615-91-C-000; DMSO through the University of Iowa; National Defense Science and Engineering Graduate Fellowships; Naval Training Systems Center N61339-93-M-0843; Sandia Labs AG-6076; NASA KSC NAG10-0122; MOCO, Inc.; DARPA N00014-90-J-1863; and NSF Instrumentation and Laboratory Improvement Program Grant USE-9152503.

References

Alibali, M. and Goldin-Meadow, S. (1993). Modeling learning using evidence from speech and gesture. In *Proceedings of the Annual Conference of the Cognitive Science Society*, Boulder. Cognitive Science Society.

Becket, W. M. (1994). The *jack lisp api*. Technical Report MS-CIS-94-01/Graphics Lab 59, University of Pennsylvania.

Cassell, J. and McNeill, D. (1991). Non-verbal imagery and the poetics of prose. *Poetics Today*, 12(3):375–404.

Cassell, J., McNeill, D., and McCullough, K.-E. (1993). Kids, don't try this at home: Experimental mismatches of speech and gesture. presented at the International Communication Association annual meeting.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: Rule based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *SIGGRAPH'94*.

Feiner, S. and McKeown, K. (1991). Automating the generation of coordinated multimedia explanations. *IEEE Computer*, 24(10).

Grosz, B. and Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3).

Hajičová, E. and Sgall, P. (1988). Topic and focus of a sentence and the patterning of a text. In Petofi, J., editor, *Text and Discourse Constitution*. De Gruyter, Berlin.

Halliday, M. (1967). *Intonation and Grammar in British English*. Mouton, The Hague.

Houghton, G. and Pearson, M. (1988). The production of spoken dialogue. In Zock, M. and Sabah, G., editors, *Advances in Natural Language Generation: An Interdisciplinary Perspective, Vol. 1*. Pinter Publishers, London.

Kendon, A. (1974). Movement coordination in social interaction: some examples described. In Weitz, editor, *Nonverbal Communication.* Oxford University Press.

Kendon, A. (1994). Do gestures communicate: A review. *Research on Language and Social Interaction.*

Liberman, M. and Buchsbaum, A. L. (1985). Structure and usage of current Bell Labs text to speech programs. Technical Memorandum TM 11225-850731-11, AT&T Bell Laboratories.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago.

Pelachaud, C., Badler, N., and Steedman, M. (1991). Linguistic issues in facial animation. In Magnenat-Thalmann, N. and Thalmann, D., editors, *Computer Animation '91*, pages 15–30. Springer-Verlag.

Power, R. (1977). The organisation of purposeful dialogues. *Linguistics.*

Prevost, S. and Steedman, M. (1993a). Generating contextually appropriate intonation. In *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pages 332–340, Utrecht.

Prevost, S. and Steedman, M. (1993b). Using context to specify intonation in speech synthesis. In *Proceedings of the 3rd European Conference of Speech Communication and Technology (EuroSpeech)*, pages 2103–2106, Berlin.

Prince, E. F. (1992). The ZPG letter: Subjects, definiteness and information status. In Thompson, S. and Mann, W., editors, *Discourse description: diverse analyses of a fund raising text*, pages 295–325. John Benjamins B.V.

Steedman, M. (1991). Structure and intonation. *Language*, pages 260–296.

Tuite, K. (1993). The production of gesture. *Semiotica*, 93(1/2).

Wahlster, W., André, E., Graf, W., and Rist, T. (1991). Designing illustrated texts. In *Proceedings of the 5th EACL*, pages 8–14.